稳健回归分析

胡良平[1,2*]

(1. 军事科学院研究生院,北京 100850;
2. 世界中医药学会联合会临床科研统计学专业委员会,北京 100029
* 通信作者:胡良平,E-mail:lphu812@sina.com)

【摘要】 本文目的是介绍稳健回归分析的概念、作用以及用软件实现计算的方法。先介绍有关的基本概念,再介绍基本原理,最后通过两个实例并基于 SAS 软件演示如何实施此分析方法。结果表明:①此法不能解决资料中存在多重共线性问题带来的坏影响;②此法能够很好地解决资料中存在异常点的回归分析问题。

【关键词】 稳健回归分析;多重共线性;M 估计;MM 估计;S 估计;LTS 估计
中图分类号:R195.1 文献标识码:A doi:10.11886/j.issn.1007-3256.2018.03.003

Robust regression analysis

Hu Liangping[1,2*]

(1. *Graduate School, Academy of Military Sciences PLA China, Beijing* 100850, *China*;
2. *Specialty Committee of Clinical Scientific Research Statistics of World Federation of Chinese Medicine Societies, Beijing* 100029, *China*
* *Corresponding author: Hu Liangping, E-mail: lphu812@sina.com*)

【**Abstract**】 The purpose of this paper was to introduce the concepts and functions and the calculation methods by using the statistical software of the robust regression analysis. Firstly, the basic concepts of the regression analysis was introduced. Secondly, the basic principle of the regression analysis was given. Finally, the robust regression analysis was demonstrated through two examples by using the SAS software. The following two distinct conclusions could be drawn: ① The method mentioned above could not solve the problem of the regression analysis of the data with the multiple collinearity. ② The method mentioned above could perfectly solve the problem of the regression analysis when there are some outliers in the data.

【**Keywords**】 Robust regression analysis; Multiple collinearity; M-estimation; MM-estimation; S-estimation; LTS-estimation

1 概 述

1.1 何为稳健回归分析

设用 Y 代表因变量,X_1、X_2、…、X_m 分别代表 m 个自变量,则多重线性回归模型可以表示为:

$$Y=\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_mX_m+\varepsilon \quad (1)$$

式中 β_0 为总体截距,β_1、β_2、…、β_m 分别为各个自变量所对应的总体偏回归系数,ε 为随机误差,常假定其服从正态分布。偏回归系数 β_i($i=1,2,\cdots,m$)表示在其他自变量固定不变的情况下,X_i 每改变一个测量单位时所引起的因变量 Y 的平均改变量。多重线性回归模型的样本回归方程可以表示为:

$$\hat{Y}=b_0+b_1X_1+b_1X_2+\cdots+b_mX_m \quad (2)$$

这里 $\hat{Y}$ 表示 Y 的估计值,b_0、b_1、b_2、…、b_m 为截距和偏回归系数的样本值,它们是相应总体参数的估计值。

如何求出模型(1)中的参数(包括截距项和回归系数)呢?当资料满足一些前提条件(例如模型的误差项服从正态分布、自变量互相独立、不存在严重的异常点)时,只需要采取普通的最小二乘法(简称 OLS 估计法,也叫做最小平方法)来构造求解回归系数的正规方程组,然后解此方程组,就可获得全部参数的估计值。但是,当资料中存在严重异常点时,就需要采用“稳健回归分析方法”来给出参数的估计值。

所谓稳健回归分析,就是在构建多重线性回归模型(1)时,不以普通的最小二乘法来构造求解回归系数的正规方程组,而是依据某些改进的做法来构造求解回归系数的正规方程组。其目的就是依据推导出来的用于估计回归参数的计算公式使参数的估计结果具有尽可能好的稳定性,即尽可能降低或消除异常点对回归分析结果的影响。

1.2 稳健回归分析应用的场合

当拟做多重线性回归分析的原始数据存在较大比例的“异常点”且自变量间不存在严重多重共线性时,采用此方法构建多重线性回归模型,可以最大限度地消除“异常点”对建模结果造成的影响。

1.3 稳健回归分析的原理

此法的关键在于使估计出来的回归系数比较稳

项目基金:国家高技术研究发展计划课题资助(2015AA020102)

定,其实质就是设法修改“普通最小二乘法”,使构造出来的正规方程组对“异常点”不敏感,再通过类似于“迭代再加权最小二乘法”等方法求解正规方程组,从而获得各回归系数相对稳定的估计值。具体的方法有多种,例如 L 估计、R 估计、M 估计、S 估计和 MM 估计等[1-2]。其中有些估计方法还可以作进一步细分,例如 M 估计可进一步分为“Huber 估计”“Tukey 估计”和“中位数估计”。由于这些估计方法涉及很深的数学知识,在文献[1]中用了 8 页篇幅介绍了前述提及的估计方法,感兴趣的读者可参阅有关文献,故此处从略。

2 基于稳健回归分析解决实际问题

2.1 此法可否解决多重共线性问题

2.1.1 问题与数据结构

【例 1】沿用文献[3]中的“问题与数据”,并基于派生变量得到的“最优回归模型”所决定的“数据集”,来提出下面的“新问题”:即“weight”的回归系数为“-88.00801”,这个“负值”表明:体重越重的人收缩压(SBP)越低,这似乎不符合临床专业知识。尽管计算出来的因变量的预测值在专业上都成立,且模型残差的方差=122.32418、R^2=0.9931,这些结果都提示所构建的多重线性回归模型很好。但毕竟存在回归系数的正负号不符合专业知识的“严重瑕疵”,这是一个需要彻底解决的“疑难问题”!

2.1.2 所需要的 SAS 程序

尝试采用“稳健回归分析方法”解决上述提及的“疑难问题”。所需要的 SAS 程序如下:

```
data a1;
  input id age height weight bmi sbp;
cards;
(此处输入文献[3]表 1 中 50 行 6 列数据)
;
run;
   /* 以上程序为了创建数据集 a1 */
data a2;
  set a1;
x1 = age * age; x2 = age * height; x3 = age * weight;
x4 = age * bmi; x5 = height * height; x6 = height * weight;
x7 = height * bmi; x8 = weight * weight; x9 = weight * bmi;
x10 = bmi * bmi;
run;
   /* 以上程序是在数据集 a1 基础上创建数据集 a2,它增添了 10 个派生变量 */
proc reg data = a2;
model sbp = age weight x3 x6 - x10/noint;
quit;
   /* 以上程序是调用 REG 过程并基于数据集 a2 拟合文献[3]中那个‘最佳’回归模型 */
proc robustreg data = a2 method = m seed = 100;
model sbp = age weight x3 x6 - x10/noint;
quit;
   /* 以上程序是调用 ROBUSTREG 过程并基于数据集 a2 和 M 估计方法拟合文献[3]中那个‘最佳’回归模型 */
```

/* 接下去,将上面 SAS 过程步中的关键词“method = m”依次修改为:method = lts、method = s 和 method = mm,就是调用 ROBUSTREG 过程并基于数据集 a2 且分别用 LTS 估计法、S 估计法和 MM 估计法来拟合文献[3]中那个“最佳”回归模型 */

【SAS 程序说明】在以上的 SAS 程序中,用“/* …… */”注释语句作了说明。

2.1.3 SAS 输出结果及其解释

以下将上面 SAS 程序中的 5 个过程步的输出结果以浓缩的方式呈现出来,见表 1。

【说明】比较上述由“REG 过程”与基于“ROBUSTREG 过程”并分别采用“M 估计法”“LTS 估计法”“S 估计法”和“MM 估计法”对同一个具有严重共线性问题的资料进行多重线性回归分析的结果可知:它们估计的“参数值”比较接近,但仍没有消除多重共线性对回归系数的严重影响(尤其是 weight 前的回归系数为负值且其绝对值还比较大,不符合临床专业知识)。也就是说:SAS/STAT 中的“ROBUSTREG 过程”不能解决“多重共线性问题”。要想消除自变量间多重共线性的影响,常用的方法有两种,第一种就是采用“主成分回归分析”,见文献[3];第二种就是采用“岭回归分析”,参见本期中的《岭回归分析》一文。

2.2 此法可否解决资料中有较多异常点问题

2.2.1 问题与数据结构

【例 2】假定有一个总样本含量 n=1000 的数据集中包含 10% 异常点的资料,每组数据(即每个个体)包含三个变量(x_1,x_2,y)的观测值,见表 2。

表 1 五种估计参数方法估计“例 1 资料”的主要结果

变量	C	E	C	E	C	E	C	E	C	E
	OLS		M		LTS		S		MM	
age	1.82182	0.49294	2.2281	0.4367	1.8457	–	2.2305	0.4273	2.1654	0.4437
weight	-88.00801	26.67636	-95.6493	23.6350	-54.9521	–	-95.7005	23.3398	-94.7289	24.2199
x3	-0.00971	0.00342	-0.0117	0.0030	-0.0084	–	-0.0117	0.0030	-0.0114	0.0031
x6	0.64569	0.19305	0.6992	0.1710	0.4000	–	0.6996	0.1688	0.6926	0.1752
x7	4.32456	1.30917	4.7009	1.1599	2.7574	–	4.7033	1.1452	4.6570	1.1884
x8	-0.05835	0.01884	-0.0647	0.0167	-0.0302	–	-0.0647	0.0166	-0.0640	0.0172
x9	0.78530	0.25836	0.8790	0.2289	0.4163	–	0.8794	0.2279	0.8694	0.2359
x10	-2.62458	0.87715	-2.9414	0.7771	-1.4683	–	-2.9427	0.7721	-2.9106	0.7996

注:C 与 E 分别代表“参数估计值”与“标准误”;OLS、M、LTS、S、MM 分别代表估计回归模型中参数的方法依次为普通最小二乘法、M 估计法、LTS 估计法、S 估计法和 MM 估计法;第 7 列上为空,因为 LTS 估计法给出的误差项是“标准差”,与其他方法不一致(其他为“标准误”),故未呈现出来

表 2 某资料中首尾各 10 组数据($n=1000$)

num	x_1	x_2	y	num	x_1	x_2	y
1	1.42151	1.13105	21.2009	991	-1.15836	-1.04066	100.179
2	2.00893	0.79083	21.1920	992	0.71023	-0.61811	100.056
3	1.82697	-0.02043	18.6024	993	1.04106	0.26261	100.389
4	-1.49036	-0.93768	-1.0254	994	-0.26348	-0.05647	100.137
5	-0.45912	-3.42593	-2.5518	995	-1.09514	-1.76723	99.993
6	0.39892	-2.02172	5.9973	996	1.65162	-0.83433	98.759
7	1.08865	-0.98391	13.0489	997	-0.26465	-0.73262	100.474
8	-0.48139	1.33821	11.2082	998	0.55630	-1.52099	100.534
9	-0.23384	-0.85954	6.0600	999	0.05137	0.24319	99.510
10	0.00900	-0.24376	9.6614	1000	-1.53994	0.96521	100.505

注:此表省略编号为 11 到 990 之间的 980 行数据;在全部 1000 行数据中,最后 100 行数据为“异常点”,占 10%

【特别说明】例 2 是人为构造的,它来自 SAS 9.3的 ROBUSTREG 过程中的“样例”。三个变量“x_1、x_2 和 y”没有实际的专业含义,仅为了造出一个样本含量为 1000 且含 10% 异常点的数据集。

设定 x_1 和 x_2 及测量误差 e 都是服从标准正态分布的随机变量(其均值为 0、方差为 1),前 900 个 y 的数值按下面的模型(3)计算出来;最后 100 个 y 的数值按下面的模型(4)计算出来:

$$y = 10 + 5 * x_1 + 3 * x_2 + 0.5 * e \quad (3)$$

$$y = 100 + e \quad (4)$$

比较式(3)与式(4)可知:y 的前 900 个数据中的每一个都在基数“10”基础上再加上三项并不大的数值,其平均值大约为“10 + 5 + 3 = 18”;而 y 的后 100 个数据中的每一个都在基数“100”基础上再加上一个随机误差,其平均值大约为 100。由此可知:表 2 的 1000 行数据中,对因变量 y 而言,后 100 个 y 值明显大于前 900 个 y 值,故属于“异常值”,它们所对应的那 100 行数据点就属于“异常点”了。

【问题】试拟合表 2 中 y 依赖 x_1、x_2 变化的二重线性回归模型。

2.2.2 所需要的 SAS 程序

(1)先用下面的一段 SAS 数据步程序产生表 2 中的 1000 行 3 列数据,创建数据集 aa。

```
data aa (drop = i);
  do i = 1 to 1000;
    x1 = rannor(1234);
    x2 = rannor(1234);
    e = rannor(1234);
    if i > 900 then y = 100 + e;
    else y = 10 + 5 * x1 + 3 * x2 + 0.5 * e;
    output;
```

```
  end;
run;
```

/* 以上程序产生 1000 组数据(x_1,x_2,y),其中,有 10% 的是异常值 */

(2)再用下面的 SAS 程序将数据集 aa(即表 2 中的数据)拷贝成数据集 a,然后再调用 SAS 过程进行建模(也可用前面例 1 中创建数据集 a1 的方法,此处从略)。

```
data a;
  set aa;
proc reg data = a;
  model y = x1 x2;
run;
proc robustreg data = a method = m seed = 100;
  model y = x1 x2;
run;
```

分别将上面的"method = m"后面的"m"替换成"lts""s"和"mm",就可得到另三种回归系数的稳健估计结果。

【SAS 过程步程序说明】第 1 个过程步调用 REG 过程创建二重线性回归模型;从第 2 到第 5 个过程步都是调用 ROBUSTREG 过程构建二重线性回归模型,其区别在于它们分别采用 M 估计、LTS 估计、S 估计和 MM 估计方法来估计模型中的参数值。

2.2.3 SAS 输出结果及其解释

以上 SAS 程序中 5 个 SAS 过程步输出的主要结果列入表 3 中。

表 3 五种估计参数方法估计"例 2 资料"的主要结果

变量	C	E	C	E	C	E	C	E	C	E
	OLS		M		LTS		S		MM	
截距	19.06712	0.86322	10.0024	0.0174	10.0083	–	10.0055	0.0180	10.0035	0.0176
x_1	3.55485	0.86892	5.0077	0.0175	5.0316	–	5.0096	0.0182	5.0085	0.0178
x_2	2.12341	0.83039	3.0161	0.0167	3.0396	–	3.0210	0.0172	3.0181	0.0168

注:C 与 E 分别代表"参数估计值"与"标准误";OLS、M、LTS、S、MM 分别代表估计回归模型中参数的方法依次为"普通最小二乘法"、M 估计法、LTS 估计法、S 估计法和 MM 估计法;第 7 列上为空,因为 LTS 估计法给出的误差项是"标准差",与其他方法不一致(其他为"标准误"),故未呈现出来

由表 3 可知:当资料中存在 10% 异常点时,基于普通最小二乘法(OLS)给出的参数估计值偏离其真值(截距 = 10、x_1 前的斜率 = 5、x_2 前的斜率 = 3)很远,而基于"M 估计法""LTS 估计法""S 估计法"和"MM 估计法"给出的参数估计值都与其真值非常接近。再列出这 5 种算法对应的复相关系数平方[即 y 与(x_1 和 x_2)的相关系数的平方]分别为:0.0234(OLS 法)、0.7788(M 估计法)、0.9932(LTS 估计法)、0.9928(S 估计法)和 0.7520(MM 估计法)。其中,OLS 估计法的复相关系数平方非常小,而 LTS 估计法的复相关系数平方最大。

2.3 小结

用 ROBUSTREG 过程时应选择哪一种"稳健估计"方法创建多重线性回归模型很难一概而论。一般来说,在拟合优度"Goodnes – of – Fit"评价的四个指标中,R – Square 取值越大、另三个评价指标(因篇幅所限,在本文中省略了)取值越小,并且,回归系数的"标准误"越小越好,该估计方法的拟合效果就越好。就本例而言,总体来看,LTS 估计法的效果最好。

最关键的问题在于:本例中的数据是基于上面的二重线性回归模型(3)产生出来的。这个模型意味着:截距项为 10、x_1 前的回归系数为 5、x_2 前的回归系数为 3,在此基础上,加一个随机误差的二分之一。此处的"随机误差"是服从均值为 0、方差为 1 的正态分布的随机误差。基于表 3 中的计算结果,可以写出五个模型的具体表达式,见式(5) – 式(9)。

$$\hat{y} = 19.06712 + 3.55485x_1 + 2.12341x_2 \quad (5)(\text{OLS 估计法})$$

$$\hat{y} = 10.0024 + 5.0077x_1 + 3.0161x_2 \quad (6)(\text{M 估计法})$$

$$\hat{y} = 10.0083 + 5.0316x_1 + 3.0396x_2 \quad (7)(\text{LTS 估计法})$$

$$\hat{y} = 10.0055 + 5.0096x_1 + 3.0210x_2 \quad (8)(\text{S 估计法})$$

$$\hat{y} = 10.0035 + 5.0085x_1 + 3.0181x_2 \quad (9)(\text{MM 估计法})$$

【结论】以模型(3)为"金标准",模型(5)偏离很远;模型(6) – (9)的质量都很高。

若再结合"复相关系数平方"等评价指标综合评价,就本例而言,LTS 估计法所得的结果最稳健。

参考文献

[1] 胡良平, 胡纯严, 鲍晓蕾. 应用数理统计[M]. 北京: 电子工业出版社, 2015: 145 – 152.

[2] 茆诗松. 统计手册[M]. 北京: 科学出版社, 2006: 497 – 510.

[3] 胡良平. 主成分分析应用(Ⅰ)——主成分回归分析[J]. 四川精神卫生, 2018, 31(2): 128 – 132.

(收稿日期:2018 – 05 – 03)

(本文编辑:唐雪莉)